Class 2: Quantifying variation in rule theories

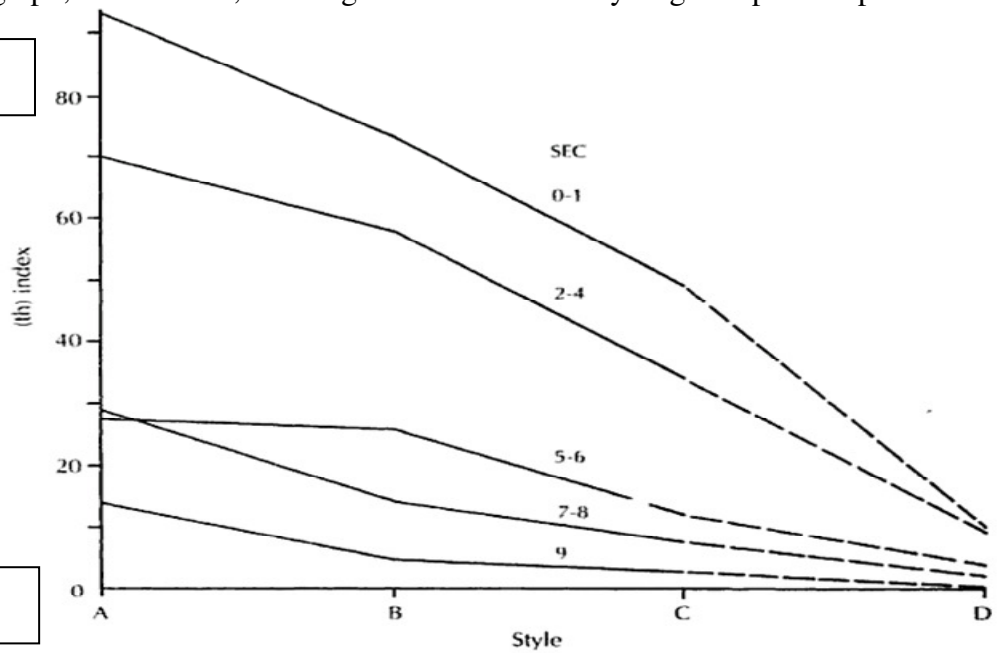
To do for Thursday
- OPTIONAL: read Cedergren & Sankoff 1974
- You might want to start on Friday's (required) reading, Anttila 1997
- Homework to prepare for in-class lab
  - Install the free software called R on a computer that you have access to (http://www.r-project.org/)
  - Go through D-Y Kim's R tutorial at http://math.illinoisstate.edu/dhkim/rstuff/rtutor.html.
  - Don't worry if you don't understand everything. Doing the tutorial will help you to get comfortable with the software; we'll go through all the commands you'll need in class.

Overview: Variable rules; linear regression as a way of modeling variation; why linear regression is not good for binary data; logistic regression as a way of modeling binary variation.

1 Free variation as stylistic variation

Classic work in sociolinguistics focused on how "variable rules" are affected by social factors. Famous graph, Labov 1972, showing how New York City English speakers pronounced /θ/:

Lots of [t̪]



Always [θ]

Fig. 4.1. Class stratification of a linguistic variable with stable social significance: (th) in *thing, through,* etc. Socioeconomic class scale: 0–1, lower class; 2–4, working class; 5–6, 7–8, lower middle class; 9, upper middle class. A, casual speech; B, careful speech; C, reading style; D, word lists.

p. 113

Labov's early approach

/θ/ → [–continuant] , optional rule

rate of applying rule = $a + b$*Class + c*Style

- Different people have different baseline rates of applying rule (a+b*Class)
- But they vary the same way in response to "style" (c*Style, where Style A=0, Style B=1, etc.)

2 Linear regression for Labov's data pattern

First, a bit more detail about Labov's **dependent variable**—the thing he's trying to predict.

- For each /θ/ in a person's recorded speech, give 1 point for [θ], 2 for [t̪θ], 3 for [t̪].
- Average these scores for that person: 3, 1, 1, 2, 1, 1, 2 → 11/7 = 1.57
- Multiply by 100: 1.57 → 157
- Subtract 100: 157 → 57

The resulting "(th) index" is a number that can range from 0 to 200.

Back to our equation:

- rate of applying rule = $a + b$*Class + c*Style

This is a **linear regression model**

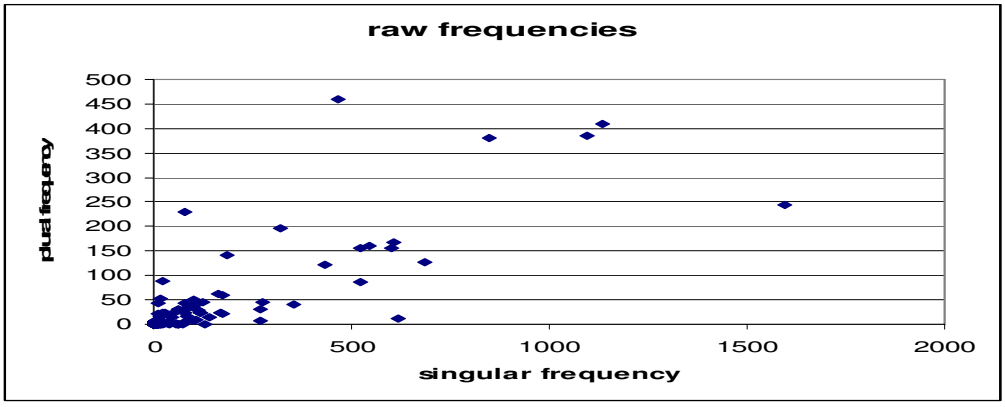
- "linear" because $y = a + bx$ is the equation for a line.
- So is $y = a + bx_1 + cx_2 + dx_3 + ex_4 + ...$

○ Let's try different values for *a*, *b*, and *c* (switch to computer projection).

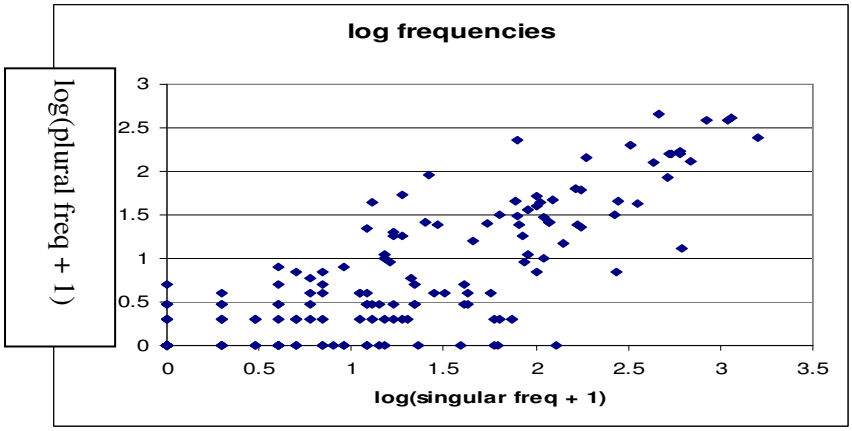
3 Fitting a regression model—introducing another set of data

- So how do we choose the best values for *a*, *b*, and *c*?
  - Let's use a different, simpler set of data.
  - Here are some data I was recently compiling from CELEX (Baayen & al. 1995): frequency of certain singular nouns vs. the frequency of their plurals.

- If we plot raw frequencies, it's difficult to read because there are a few items with very high frequencies, so most items are crowded in a corner.
  - This is typical of word frequencies:



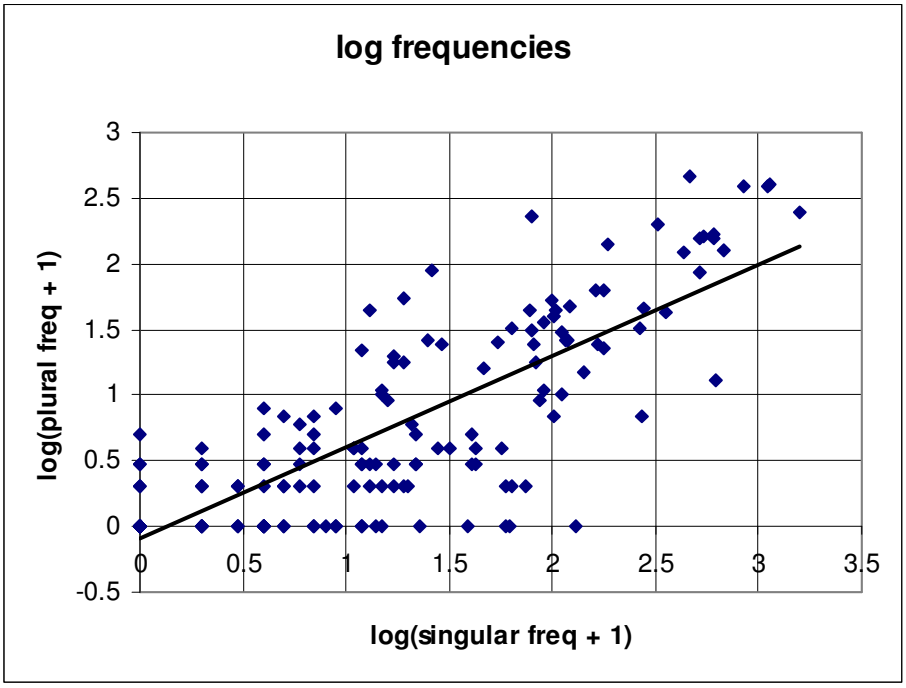
- So, instead of raw frequencies, we plot the **log** of the frequency (next page)



Math review: can you guess why I used log(freq + 1) instead of log(freq)?

4 Fitting a regression model—the regression line

- The data aren't in a perfect straight line, but they're sort of grouped around a line.
  - Here's a line that seems to go through the middle of the points:



Math review: What is the equation of this line (approximately)?

- This line is a **regression line**—it represents a **regression model**.
  - $y = a + bx$ a (intercept) and b (slope) are **coefficients** in the regression model.

- MS Excel chose this line for me. Why did it choose *this* line?
  - We need some measure of how well the line **fits** the data.
  - The measure typically used in linear regression is called **least squares**:
    - Minimize the sum of the squared vertical distances between the points and the line (that is, the difference between actual y-value and y-value on the regression line)
- Let's look at an animation from Yihui Xie's wiki at http://animation.yihui.name/lm:least_squares (switch to projector)

Summary: the computer finds the values for a and b that minimize the sum of squared distances between y and $a+bx$.

We won't get into *how* the computer finds those values (can be done with matrix algebra or by "hill-climbing", as in the animation)—you don't need to know that unless you're designing statistics software.

5 Significance

- Here's our regression line equation for singular-vs-plural
  - log(plural + 1) = -0.0886 + 0.6915 * log(singular + 1)
- We're saying that the plural's log frequency is about 70% as large as the singular's log frequency.
  - Seems to claim that the plural frequency tends to *depend* on the singular frequency.
  - Is this really true? Or is the relationship random?
- One way to answer that question is to ask
  - *How **surprising** would* the *pattern be if there were no relation between singular and plural frequency?*

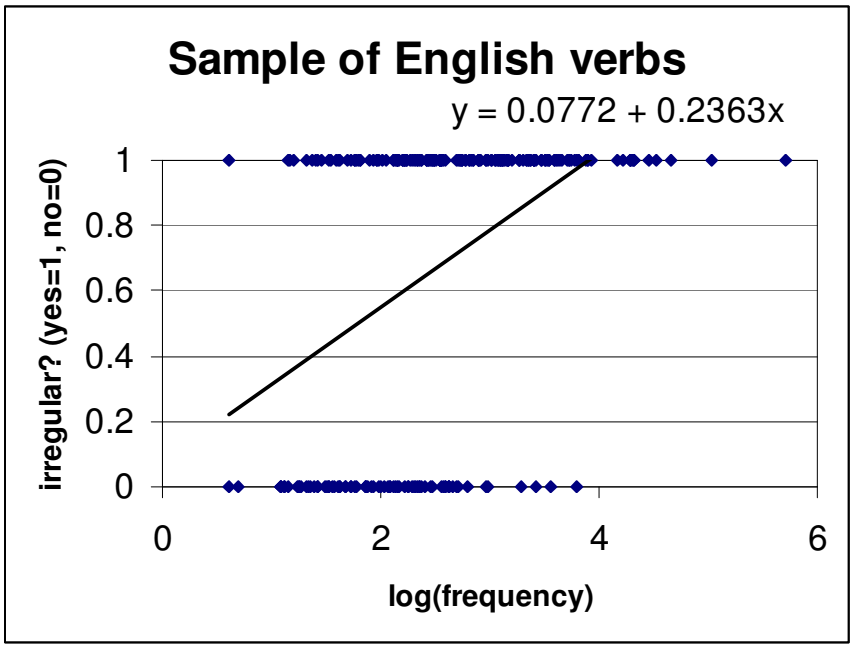
- or, *If we generate fake data with no relation between singular and plural frequency 100 times, how often would we see a pattern this strong?*
- We don't have to actually generate fake data in this case. Because of some convenient mathematical properties of linear regression models, the computer can estimate for us.
  - In this case, a coefficient as big as 0.6915 would happen by chance less than 1 in 10^{16} times.
  - Or, in statistics-speak, $\boldsymbol{p} < 0.0000000000000001$

We will play with linear regressions in the in-class lab on Thursday.

6 Binary dependent variables

- What if our **dependent variable** is not a number (like the (th) index that ranges from 0 to 200), but instead "yes" or "no"?
  - As a first try, let's set "yes"=1 and "no"=0, and treat it as a number.

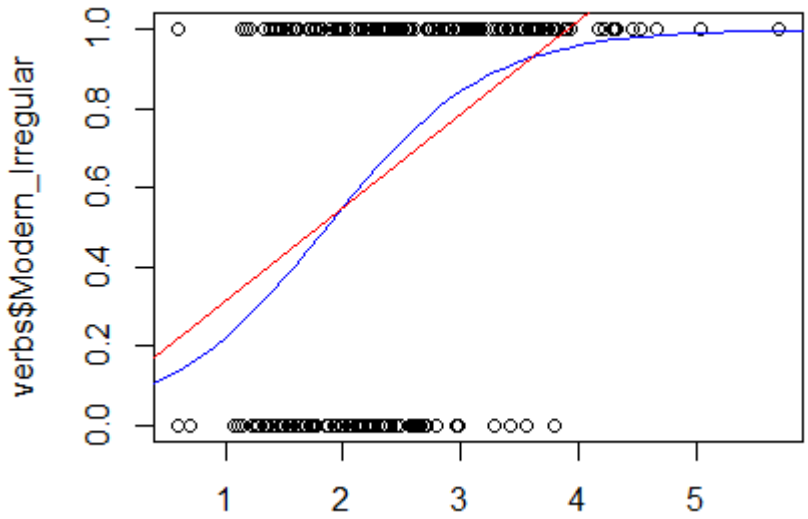
Data from Lieberman & al 2007:
- for English verbs that existed in Old English, are they now regular (0) or irregular (1)?
- How well does (modern) frequency predict it?



- This is not quite right though.
  - For a word with frequency 100, what is the predicted value of *y*? What does that mean??
  - If we ask the computer to estimate *p*-values for us (how often is such a large intercept or slope expected by chance), results will be inaccurate
    - This is because the estimate requires the points to be a similar distance from the line across the range of the *x* value (frequency).
    - In this case, the (unreliable!) *p*-value for the slope is $p = 0.000000000000174$

7 Logistic regression

- Instead of trying to fit a line, we ask the computer to fit an s-shaped curve:



The computer's job was to find these two numbers (coefficients)

- The equation for this curve: $y = \frac{1}{1+e^{-(-2.73+1.47*x)}}$
  - Let's play with some different values of those two numbers and see what happens (change to projector)

- This type of function, $y = \frac{1}{1+e^{-(a+b*x_1+c*x_2+d*x_3+\ldots)}}$, is known as a logistic function.
  - Sometimes called "sigmoid" because it's shaped like "S" (from Greek letter name "sigma"), though there are other functions that also could be called sigmoid.

- How did the computer find the coefficients?
  - By hill-climbing (as in our animation before)—it keeps adjusting until the fit stops improving.

- Significance:
  - So does frequency affect whether a verb is irregular?
  - That is, is the slope coefficient 1.47 surprisingly large?
  - The computer tells us that if frequency and irregularity had no relation, we would get such a large slope only 0.0000000201% of the time, or $p = 0.000000000201$

8 Back to sociolinguistics: Varbrul

- Eventually, sociolinguistics researchers settled on using logistic regression, sometimes called Varbrul (variable rule) analysis.
- Various researchers, especially David Sankoff, developed software called GoldVarb (Sankoff & al. 2005 for most recent version) for doing logistic regression in sociolinguistics.
  - Goldvarb uses slightly different terminology though.
  - If you're reading sociolinguistics work in the Varbrul, see Johnson 2009 for a helpful explanation of how the terminology differs.

If we have time: let's try adding more **independent variables** to our English-verb regression model.

Next time:
- First, a lab where we practice observed/expected tables, linear regression, and logistic regression (including with multiple factors).
- Second, model comparison

References

Baayen, R.H., R. Piepenbrock & L. Gulikers. 1995. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and Linguistics Compass*. 3(1): 359-383.

Labov, William. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449: 713-716.

Sankoff, David, Sali Tagliamonte, and Eric Smith. 2005. GoldVarb X: a variable rule application for Macintosh and Windows. <http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm>